文章编号: 1002-0411(2001) 07-0676-05

一种基于马氏距离建立简化多元统计模型的方法

高 翔[1] 王 纲[2] 马纪虎[1]
(1. 中国科学院沈阳自动化研究所 沈阳 110015; 2. 沈阳化工学院 沈阳 110021)

摘 要: 提出一种基于样本之间最小马氏距离的样本平均方法, 从总体正常历史采样数据样本集合中, 构造新的数据样本集, 建立简化多元统计模型. 然后通过判断两数据集的质心偏移和协方差的差异程度来检验新的数据样本集对总体样本集的可代表性, 从而达到用较少的有效样本代表总体样本统计特征的目的. 仿真结果表明用本文提出的简化多元统计模型进行故障诊断的效果与传统模型相同, 而降低了对系统存储量和计算量的要求.*

关键词: 马氏距离; 主元分析; 多元统计模型; 多元校验; 可代表性; 故障诊断
中图分类号: TP206 文献标识码: B

AN APPROACH OF BUILDING SIMPLIFIED MULTIVARIATE STATISTICAL MODEL BASED ON MAHALANOBIS DISTANCE

GAO Xiang[1] WANG Gang[2] MA Ji-hu[1]
(1. *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110015 *China*;
2. *Shenyang Institute of Chemical Technology, Shenyang* 110021 *China*)

Abstract: An averaging sample approach based on the shortest Mahalanobis Distances (MD) among the samples is proposed. A new data set of samples is constructed from the data set of total normal historical samples for building multivariate statistical models. The representativity of the data set of new samples related to the total samples can be examined through comparing the deviation of centroids and difference of the covariance between the two data sets so that a goal is achieved which the statistical characteristic of total samples is represented by only using fewer effective samples from all samples. Simulation result proves the sameness between the simplified model and the previous model for fault diagnosis so that there are lower requirement for memory and computation in the system.

Keywords: mahalanobis distance, principal component analysis, multivariate statistical model, multivariate calibration, representativity, fault diagnosis

1 引言(Introduction)

多元统计过程控制(Multivariate Statistical Process Control, MSPC) 是近年来发展起来的一种过程监控和故障诊断方法, 通过采集生产过程中大量历史数据, 从中选取处于正常操作条件(Normal Operation Condition, NOC) 下的样本建立多元统计模型, 进行实时多元统计分析, 及时发现过程的异常变化[1].

多元统计分析的一个基本假设是: 系统中所有处于 NOC 内的数据样本, 都符合多元统计模型历史数据集合的统计规律. 所以, 正确选择统计模型的样本, 使之能够准确地反映系统的异常情况, 是准确进行故障诊断的前提和基础. 如何在已采集的海量的数据总体样本集中选择一部分样本集用来建模, 而该样本集能基本上表示正常的总体样本的统计特征, 是我们面临的一个实际问题. 如果任意选取样本, 则有可能出现选中的部分样本没有涵盖总体样本范围的情况, 不可避免地给建立历史数据统计模型带来误差. 目前未见关于使用精简样本建立多元统计模型的报道.

马氏距离(Mahalanobis Distance)[2] 给出了空间中样本之间的加权距离的定义. 通过研究样本之间的马氏距离, 本文提出基于样本之间最小马氏距离的样本平均方法合理地选取样本, 可以用于建立简

* 收稿日期: 2001- 05- 12

化的历史数据模型,该模型能够代替用总体样本建立的历史数据模型成功地进行故障诊断;而且,基于马氏距离计算两数据集的质心的偏移程度和协方差差异程度[3]可以判断该模型与总体样本模型的近似程度.比较简化多元统计模型和未简化的多元统计模型,用主元分析法(Principal Component Analysis, PCA)[1]对过程进行故障诊断的结果也是相同的.

2　PCA 和 MPCA 原理简介(Brief introduction to principle of PCA and MPCA)

主元分析法是多元统计分析中最基本的方法之一.设数据矩阵 $X(m\times n)$,即 m 个历史正常数据样本,n 维原始过程变量,为了避免过程变量不同量纲对计算结果的影响,对建模数据进行量化处理后,分解成得分矩阵(Score Matrix) $T(m\times a)$ 和载荷矩阵(Loading Matrix) $P(n\times a)$ 的乘积,提取原始变量间的关联信息.在 T 和 P 内部,各个列向量之间(t_i 和 t_j; p_i 和 p_j; $i, j = 1, 2, \cdots, n$)都满足相互内积为 0,自身长度为 1 的条件.而前 $a(a< n)$ 个 t_i 即可描述原始过程变量的绝大部分信息.这样,可以用主元分析法对生产过程进行监控和故障检测等.

间歇过程是一个有限操作周期的生产过程,所以每次间歇过程数据集合构成了一族过程变量的时间轨迹,其历史数据集合为三维矩阵,可以表示为 $\underline{X}(I\times J\times K)$.其中 I 代表批次,J 代表过程变量,K 代表时间序列.

多向主元分析法(Multiway Principal Component Analysis, MPCA, Nomikos and MacGregor[4])将三维数据矩阵 $\underline{X}(I\times J\times K)$ 展开,形成一个新的两维矩阵 $\underline{X}(I\times JK)$.

在 MPCA 基础上,最小窗口多向主元分析法(Minimum Window Multiway Principal Component Analysis, MWMPCA)[5]类似 MPCA 的展开方法,将三维数据矩阵展开,形成 K 个相互独立的 $\underline{X}(I\times J)$ 矩阵,各自建立 K 个子 PCA 模型,其后的处理方法与 PCA 相同.

3　马氏距离 (Mahalanobis Distance)

在原始变量空间中,马氏矩离(Mahalanobis Distance, MD)[2]是考虑样本中变量间相关性的各样本至样本平均值的距离;而欧氏距离(Euclidean Distance, ED)只是各样本至样本平均值的几何距离.其中,用主元分析法计算的各样本的欧氏距离和马氏距离为 (t_i, t_j 为得分矩阵 T 的行向量):

$$ED_i = \sqrt{t_i t_i^{\mathrm{T}}} \quad (1\leqslant i\leqslant m) \tag{1}$$

$$MD_i = \sqrt{t_i \mathrm{cov}(T)^{-1} t_i^{\mathrm{T}}} \quad (1\leqslant i\leqslant m) \tag{2}$$

$$\mathrm{cov}(T) = \frac{1}{m-1} T^{\mathrm{T}} T \tag{3}$$

两样本之间用主元分析法计算的马氏距离为:

$$MD_{i,j} = \sqrt{(t_i - t_j)\mathrm{cov}(T))^{-1}(t_s - t_f)^{\mathrm{T}}} \quad (1\leqslant i, j\leqslant m) \tag{4}$$

其中,$\mathrm{cov}(T)$ 为 X 的得分矩阵 T 的协方差矩阵;t_i 为 T 的第 i 个行向量;

文[2]证明了在选取所有主元(个数为 n)的情况下,$MD_i^t = MD_i^o$(用原始变量计算的马氏距离).在原始变量个数特别多的情况下,选取全部主元会造成计算上的困难,用方差贡献率法选取主元时,要求所选特征值贡献率达到99%,MD_i^t 和 MD_i^o 即可基本相等.

4　两个统计模型的相互可代表性(The representativity of two statistical models)

在多元统计分析过程中,用不同的数据模型投影计算得到的计算结果是不同的,但是,我们可以比较两数据集协方差的差异程度和质心(centroids,即各样本的平均值)的偏移程度[3],从而判断两个数据集相互之间是否有可代表性.

4.1　两数据集协方差的差异程度

通过比较两数据集的协方差结构,可以比较它们的方向和分散程度的相似性:如它们的协方差矩阵不同,意味着两数据集方向上的不同,或在同一范围内变化程度的不同,抑或二者皆不同.两个协方差矩阵的等同与否也是实行 Hotelling T^2 测试的必要条件,它们的比较基于 Bartlett 测试的一般性.用 PCA 方法,$X_1(m_1\times n)$, $X_2(m_2\times n)$ 共有的协方差矩阵 $\mathrm{cov}(X_{1,2})$ 为:

$$\mathrm{cov}(X_{1,2}) = [(m_1 - 1)\mathrm{cov}(X_1) + (m_1 - 1)\mathrm{cov}(X_2)]/(m_1 + m_2 - 1) \tag{5}$$

Hotelling T^2 统计量常用来检验过程系统新的采样值是否落在主元确定的控制域内,它由下式定义:

$$T^2 = t\mathrm{cov}(T)^{-1} t^{\mathrm{T}} \tag{6}$$

与式(2)相比,显而易见,Hotelling T^2 统计量即为马氏距离的平方.

$\mathrm{cov}(X_{1,2})$ 具有 $(m_1 + m_2 - 2)$ 的自由度,用 Bartlett 法计算数值 C

$$C = (1/v)\{(m_1 + m_2 - 2)\ln(\hat{u}\mathrm{cov}(X_{1,2})\hat{u}) - (m_1 - 1)\ln(\hat{u}\mathrm{cov}(X_1)\hat{u}) - (m_2 - I)\ln(\hat{u}\mathrm{cov}(X_2)\hat{u})\} \quad (7)$$

这里,

$$v = 1 + \frac{2n^2 + 3n - 1}{6(n+1)}\left(\frac{1}{m_1 - 1} + \frac{1}{m_2 - 1} - \frac{1}{m_1 + m_2 - 2}\right) \quad (8)$$

而

$$C_{\mathrm{crit}} = x^2(n(n+1)/2) \quad (9)$$

如果, $C \leqslant C_{\mathrm{crit}}$, 则两数据集协方差结构基本等同, 否则, 就有所区别.

4.2 比较两数据集质心之间的马氏距离

要检验两数据集在空间内是否有相似的位置, 两数据集质心之间的马氏距离是具备相似位置的基础.

用 Hotelling T^2 测试计算 $X_1(m_1 \times n)$, $X_2(m_2 \times n)$ 的之间的马氏距离的平方为:

$$MD^2_{1,2} = (\bar{x}_1 - \bar{x}_2)^{\mathrm{T}}\mathrm{cov}(X_{1,2})^{-1}(\bar{x}_1 - \bar{x}_2) \quad (10)$$

$$MD^2_{1,2\mathrm{crit}} = \frac{n(m_1 + m_2)(m_1 + m_2 - 2)}{m_1 m_s(m_1 + m_2 - n - 1)}F(m, m_1 + m_2 - n - 1) \quad (11)$$

这里, $MD^2_{1,2\mathrm{crit}}$ 为容许距离平方限, 当 $MD^2_{1,2} \leqslant MD^2_{1,2\mathrm{crit}}$ 时, 质心位置相似, 满足两数据集可代表性的充分条件; 否则, 两数据集质心之间存在显著位移, 两数据集不可以互相代表.

总而言之, 两数据集协方差差异程度和质心是否位于相似的位置是问题的关键. 如两个条件均满足, 则两数据集互有代表性.

5 构造简化模型样本的方法(The approach of building simplified model samples)

PCA 过程监视的原理是所检验的某个样本能否多元投影到所建立的模型中去, 而所建模型的范围应基本上代表 NOC 的范围. 所有构成模型的样本尽可能地反映投影域.

首先计算各样本距质心的马氏距离, 删除离质心距离过远而不在 NOC 内的样本. 设剩下的样本构成一个空间, 让我们以质心为中心, 能包容全部样本的长度为半径, 做一个超球, 构成域 Ω, 则 Ω 是凸的. 因两样本间马氏距离相近的两个样本基本上特征相似, 而取它们的平均值则是保留它们共有的特征. 该平均值在域 Ω 内, 就可以继承那两个样本而最大限度地保留域 Ω 的特征. 所以可以计算各样本之间的马氏距离, 找出其中最小距离的两个样本, 对它们进行平均, 然后用这平均值代替这两个原有的样本, 这样, 总体样本的数目便逐一减少. 以此循环下去, 直至达到要求的样本数目.

设有一个总体样本 $X(M \times n)$, 要求选出的样本 $Y(m \times n)$, 其中 $m \leqslant M$. 则其选取步骤如下:

1. $Z(M \times n) = X(M \times n)$;

2. 对 $Z(M \times n)$ 进行奇异值分解, 取主元个数为 a, 使方差贡献率达到 99% 以上;

3. 由式(2)计算各样本至质心的 MD_i^t, 设平均值为 $\overline{MD}$, 删去超过 3 $\overline{MD}$ 的样本(由 3σ 准则; $\overline{MD}$ 为距质心的标准差, 99% 的正常样本在 NOC 之内).

4. 由式(4)计算各样本之间的 MD_{ij}^t, 并置于马氏距离矩阵 MDMAT($M \times M$)中, 为了便于下一步计算 min, $MD_{i,j}^t$, MDMAT(i, i) $\leftarrow \infty$ ($1 \leqslant i \leqslant M$);

5. 找出 MDMAT 中值为最小的元素:

1) 依次找出每行中的最小值, 置于行向量 $V(1 \times M)$ 中, 且依次将各最小值在每行的位置记入行向量 $S(1 \times M)$ 中;

2) 在行向量 V 中找出最小值, 设该值在 V 中的位置为 i, 则行向量 S 第 i 个元素的值设为 j, 则(i, j)就是要找的元素; $Z(M \times n)$ 中第 i 个样本和第 j 个样本之间存在最小马氏距离.

6. 矩阵 $Z(M \times n)$ 中第 i 个样本和第 j 个样本进行平均后形成一个新的样本代替原有的两个样本, 这样, Z 的维数就降低一维.

$$Z(i,:) \leftarrow [Z(i,:) + Z(j,:)]/2; \quad (12)$$

$$Z(j,:) \leftarrow [\quad]; \quad (13)$$

式(13)将空赋予 Z 的第 j 行, 意即删去这一行.

7. 判断取平均值样本后 Z 的行的维数 $M-1$ 是否等于 m, 条件满足则结束; 否则, 转步骤 2.

6 应用实例(The example of application)

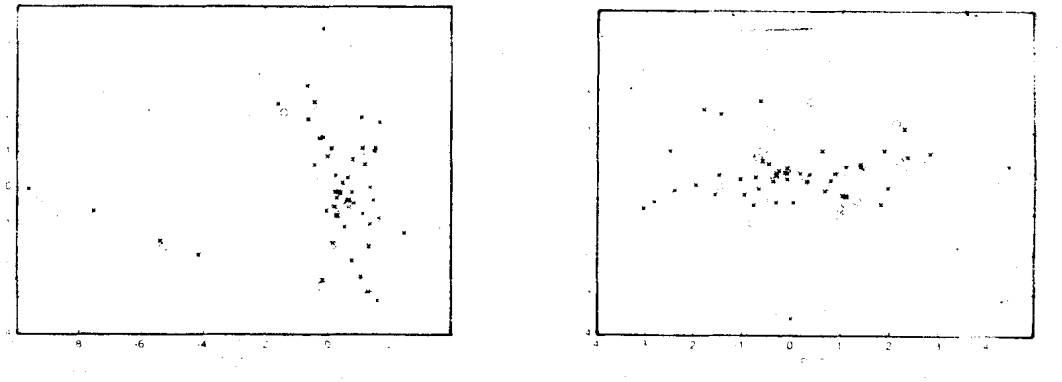
本文应用研究中的实际数据集采自某化工厂聚氯乙烯间歇过程 50 个批次, 按第 2 节给出的原则, 利用这些批次的数据集建立 MPCA 模型, 通过展开得到二维 PCA 模型, 这样 MPCA 的一个批次就相当于 PCA 中的一个样本. 现在我们假设用 50 个批次数据来建模计算量太大, 希望用 30 个批次数据来建模, 并比较其结果.

6.1　多元统计模型的简化和两模型可代表性的验证

首先, 可以用这 50 个批次数据 $\underline{X}$ (50×10×3405) 建立 MWMPCA 模型, 通过展开得到 3405 个二维 PCA 模型 X (50×10). 通过第 2 节的方法在各子 PCA 模型内求取主元, 用式(2) 和式(4) 在各 PCA 模型中分别计算出各样本的马氏距离和样本间的马氏距离. 再把各 PCA 模型之间相同序号的每个样本的马氏距离和样本间的马氏距离相加, 得到每个样本(批次) 的马氏总距离和样本间的马氏总距离. 以下样本平均的步骤可按第 5 节步骤 5, 6, 7 进行. 图 1 描述了其中一个子模型主元表示的样本分布情况.

最后, 我们需要知道新得出的 30 个样本在多元统计分析中能否代替原有的 50 个样本. 首先, 按 4.1 节, 应用式(7) 和式(8), 先计算出各子 PCA 模型的 C, 存入列向量 C_{set} (3405×1) 中去, 图 2 显示: 各子 PCA 模型的 C 值, 均小于 C_{crit} ($C_{crit}=x^2(55)$), 也就是说, 按照这种方法各 PCA 子模型简化后的协方差结构与原模型的差异程度在容许范围内, 符合可代表性的必要条件; 其次, 计算各子 PCA 简化模型与未简化模型的质心的马氏距离, 结果在各子模型中, $MD^2_{1,2}$ 均小于 $MD^2_{1,2crit}$, 也符合可代表性的充分条件.

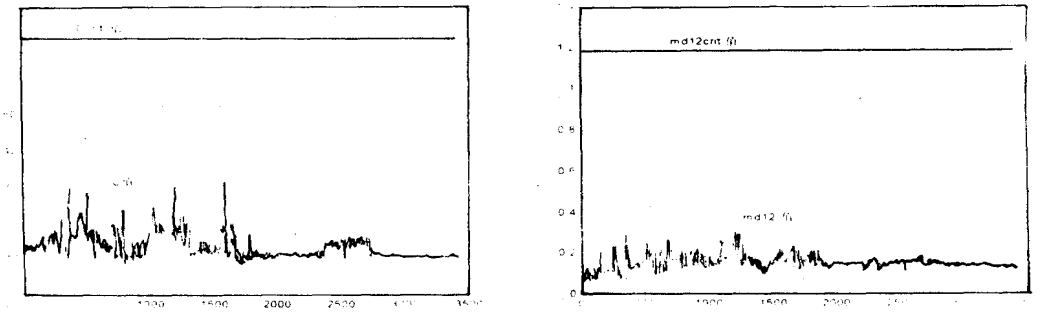
这样, 我们可以类似地在 M 个多维数据总体样本中计算出 m 个样本. 依上述判断方法, 若符合条件, 用这 m 个样本可建立简化 PCA (或 MPCA) 模型; 若不符合条件, 则说明样本数目太少, 则可计算 p ($m\leqslant p\leqslant M$) 个样本, 使之符合上述条件; 若 $p-1$ 个样本不符合条件, p 符合, 则 p 个样本所建模型称为最简化多元统计模型.



(a)　主元 $pc1$: $pc2$　　b)　主元 $pc2$: $pc3$

图 1　第一个子模型中各主元所表示的样本分布 ‘o’: 新模型样本; ‘x’: 原模型样本

Fig. 1　The distribution of some principal components in the first sub model. ‘o’: the samples in the new model. ‘x’: the samples in the previous model)



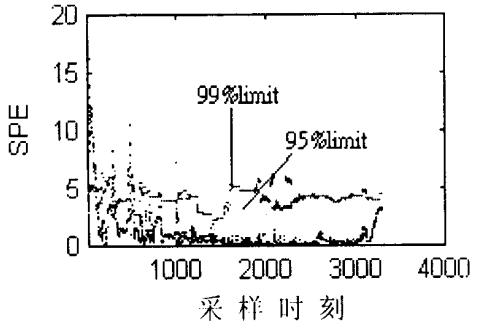
(a)　各协方差的差异　　(b)　各质心的差异

图 2　间歇过程中未简化 PCA 子模型和简化 PCA 子模型的协方差和质心的差异情况

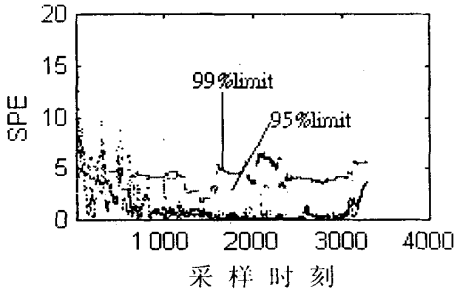
Fig. 2　The differences of the covariance and centroids between the previous sub PCA models and the simplified sub PCA models in the batch process

6.2 用简化模型进行故障诊断结果比较

平方预测误差 SPE(Square Prediction Error)[6] 的计算是统计过程控制中一种最为有效的方法,可以早期发现异常. 所以,有必要用 6.1 节建立的简化的 MWMPCA 模型和包含总体样本的 MMPCA 模型进行故障诊断,分别计算 SPE,然后进行比较. 得出的结果如图 3,从图中可以看出,过程初始段传统模型标示出 SPE 超限之处,简化模型同样在初始段可以检测出 SPE 超限,证明了简化模型的有效性.



(a) 未简化模型



(b) 简化模型

图 3 比较未简化 MPCA 模型和简化 MPCA 模型 SPE 的计算结果

Fig. 3 The comparison of the results of SPE computation with the previous and simplified MPCA models

7 结束语 (Conclusions)

用判别马氏距离的方法,找出最近样本间马氏距离的一对样本取平均值以代替这一对样本,是多元统计分析理论中的一种新型的选取样本建立简化的多元统计模型的方法. 其特点在于在规定建模样本个数的要求下,用人为计算出的部分"虚拟"样本(样本平均值)和剩余其它原有样本涵盖总体样本的空间,基本并且能够表示原数据集的统计特征. 使得应用 PCA 多元投影进行故障诊断时,不至于因建模误差产生误报和漏报. 尽管简化模型不完全等于总体样本模型,但确是总体样本模型的一种良好的逼近,而且其诊断效果不逊于总体样本模型的效果.

参 考 文 献 (References)

1 Kresta J, MacGregor J F, Marlin T B. Multivariate Statistical Monitoring of Process Operating Performance. Canadian Journal of Chemical Engineering, 1991, **69**: 35～47

2 Maesschalck De R, Jouan-Rimbaud D, Massart D L. Tutorial: The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 2000, **50**: 1～18

3 Jouan-Rimbaud D, Massart D L, Saby C A, Puel C. Characterisation of the Representativity of Selected Sets of Samples in Multivariate Calibration and Pattern Recognition. Anal. Chim. Acta 1997, **350**: 149～161

4 Nomikos P, MacGregor J F. Monitoring batch Processes using Multiway Principal Component Analysis. AIChE J. 1994, **40**: 1361～1375

5 赵立杰,王 纲. 非线性主元分析故障诊断方法及其应用. 信息与控制, **2001**(5)

6 Jackson J E. Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics 1979, **21**: 341～347

作者简介

高 翔(1967–),男,博士生. 研究领域为故障诊断、统计过程控制等.

王 纲(1956–),男,教授. 研究领域为复杂工业过程建模控制和优化、故障诊断、统计过程控制等.

马纪虎(1940–),男,研究员,博士生导师. 研究领域为复杂工业过程建模控制和优化、系统仿真、工业现场总线技术等.